社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
NC97-10 (1997-05)



# 強化学習を用いた球技動作の予測と制御

落合 真吾 佐野 雅己 沢田 康次

東北大学大学院情報科学研究科
〒980-77 仙台市青葉区片平2丁目1-1
E-mail: shingo@sawada.riec.tohoku.ac.jp

あらまし

神経回路モデルにより、動的な予測・学習・制御を行わせる。簡単な例として、ボールのシュート、壁打ちという球技動作を取り上げる。これらの問題においては一般に正確な教師信号は与えられない。ボールのシュートの場合には、入った場合にのみ「入った」という情報が与えられ、どのくらいゴールに近いかというような情報は与えられない。また、運動の制御はリアルタイムに行われるため、高速な学習アルゴリズムが要求される。本報告では、RBF を用いた予測と強化学習（TD 学習）を組み合わせてモデルを構成し、計算機シミュレーションを行った。

キーワード 確率的強化学習、TD学習、RBF

# Reinforcement learning algorithm for prediction and control of ball games

Shingo Ochiai Masaki Sano Yasuji Sawada

Graduate School of Information Sciences, TOHOKU University
2-1-1 Katahira, Aoba-ku, Sendai 980-77, JAPAN
E-mail: shingo@sawada.riec.tohoku.ac.jp

Abstract

An interesting question about learning is how an embedded agent can improve performance while acting in complex dynamical environment. Supervised learning is not feasible because precise knowledge of dynamical environment and correct response of agent are not available *a priori*. As a case study of control problem, we choose two types of ball playing games; shooting in basketball and squash tennis. In both cases, robust and efficient algorithm is needed. We applied the stochastic reinforcement learning method with temporal difference (TD) algorithm for controlling and prediction. The result was successful.

key words stochastic reinforcement learning, TD learning, RBF

## 1 はじめに

神経回路モデルにより、動的な予測・学習・制御を行わせる。簡単な例として、ボールのシュート、壁打ちという球技動作を取り上げる。これらの問題においては一般に正確な教師信号は与えられない。そのため、強化学習的な枠組で考える必要がある。

ここで扱うボールのシュートの場合には、入った場合にのみ「入った」という情報が与えられ、どのくらいゴールに近いかというような情報は与えられない。ボールがどのような運動をするかを予め知らない場合は、試行することによりアクティブな探索を行う必要がある。また、テニスの壁打ちなどの動作においては、プレーヤーは、ボールの運動法則を学習する必要があるし、ボールの落下地点を予測し、予めその位置でボールを待ちうけなければ打ち返す事ができない。そのため、外界をモデル化し、予測しながら運動を制御するアルゴリズムが必要となる。また、このような運動の制御はリアルタイムに行われるため、高速な学習アルゴリズムが要求される。すなわち、動的な学習・制御問題で要求されることは、以下のようにまとめられる。

1. 簡単な報酬信号による学習
2. アクティブな探索
3. 外界のモデル形成と予測
4. 高速な学習アルゴリズム

本報告では上記の課題に対して、(1)TD 法による強化学習 [1, 2, 3, 4]、(2) ノイズを積極的に利用した確率的強化学習、 (3) ニューラルネットによる運動予測、(4)RBF(Radial Basis Function) ネットによる学習の高速化を組み合わせてモデルを構成し、計算機シミュレーションを行いその有効性を検討する。

## 2 モデル

### 2.1 ボールのシュートのアルゴリズム

ある位置からボールをなげ、ゴールにいれる動作を考える。(図 1) ボールの運動は、粘性抵抗を無視したニュートン方程式に従い、壁や床での反射は、非弾性衝突 (跳ね返り係数 $e = -0.6$) を仮定する。ボールがゴールに入った場合に、報酬 (評価) $r$ が得られるとする。また、プレーヤーは、ボールの運動法則や壁の跳ね返り係数などの情報を知らないものとし、ボールの打ち出しの速さ $u$、角度 $\theta$ が制御できるとする。(ボールの直径は、0.4 とし、ゴールのリングの大きさを 1.0 とした。)

$$u = u_{max} \cdot g(w_u + n_u) \tag{1}$$

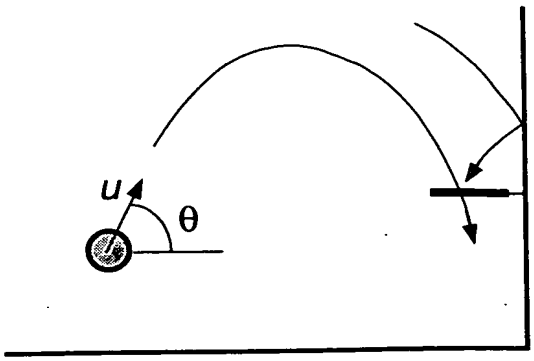


図 1: シュートの概念図

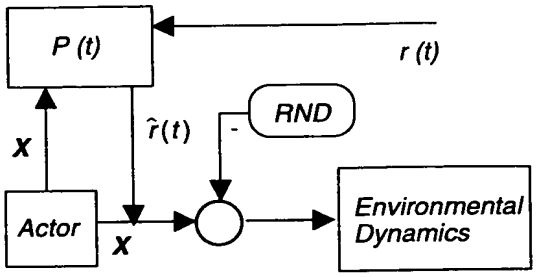


図 2: シュート制御のモデル

$$\theta = \theta_{max} \cdot \tanh(w_\theta + n_\theta) \tag{2}$$

$$\left(g(x) = \frac{1}{1+\exp(-x)}\right)$$

ここでは、ボールの速さ $u$、角度 $\theta$ を直接制御するのではなく、$w_u$、$w_\theta$ によって $u$、$\theta$ を制御する。$n_u$、$n_\theta$ はノイズであり、初期の段階ではこれにより報酬の得られる $(u, \theta)$ を探索する。

$$n_u = N_u \cdot rnd \tag{3}$$

$$n_\theta = N_\theta \cdot rnd \tag{4}$$

($rnd$ は $[-1, +1]$ の一様乱数)

近い将来に得られる報酬の和 (累積報酬) $V(t)$ を考える。

$$V(t) = \sum_{i=1}^{\infty} \gamma^{i-1} \cdot r_{t+i} \tag{5}$$

ただし、$\gamma < 1$ とする。ここで、$V(t)$ を $P(t)$ で近似する。RBF(Radial Basis Function) を用いて、

$$P(t) = \sum_{ij} v_{ij} \cdot b_{ij}(u, \theta) \tag{6}$$

とする。ここで、$v_{ij}$ : 荷重値、$b_{ij}(u, \theta)$ : basis function である。

$$b_{ij}(u, \theta) = \exp(-\frac{(u-\mu_i)^2}{2\sigma_u^2}) \cdot \exp(-\frac{(\theta-\nu_j)^2}{2\sigma_\theta^2}) \tag{7}$$

$b_{ij}(u,\theta)$ は、$i$、$j$ によって指定され、中心を $(\mu_i,\nu_j)$ とするガウス関数である。
$V(t)$ は、

$$\begin{aligned} V(t-1) &= \sum_{i=1}^{\infty}\gamma^{i-1}\cdot r_{t-1+i} \\ &= r_t+\sum_{i=2}^{\infty}\gamma^{i-1}\cdot r_{t-1+i} \\ &= r_t+\gamma\sum_{i=1}^{\infty}\gamma^{i-1}\cdot r_{t+i} \\ &= r_t+\gamma V(t) \end{aligned} \tag{8}$$

と書ける。$P(t)$ の近似が完全ならば、$P(t)$ もこの式を満たし、

$$P(t-1)=r+\gamma P(t) \tag{9}$$

となる。ここで、誤差 $\hat{r}$ を、

$$\hat{r}=r+\gamma P(t)-P(t-1) \tag{10}$$

とする。これを用いて、式 (6) の荷重値 $v_{ij}$ の変更を計算すると

$$\begin{aligned} \Delta v_{ij} &\propto -\frac{\partial\frac{1}{2}\hat{r}^2}{\partial v_{ij}}=-\hat{r}\cdot\frac{\partial\hat{r}}{\partial v_{ij}} \\ &= -\hat{r}\cdot\left[\gamma\frac{\partial P(t)}{\partial v_{ij}}-\frac{\partial P(t-1)}{\partial v_{ij}}\right] \end{aligned} \tag{11}$$

となるが、このうち後ろの $P(t-1)$ の項を用いて更新する。これは、過去の予測の方を訂正するということである。したがって、$\Delta v_{ij}$ は、

$$\begin{aligned} \Delta v_{ij} &\propto \hat{r}\cdot\frac{\partial P(t-1)}{\partial v_{ij}} \\ &= \hat{r}\cdot b_{ij}(u,\theta) \end{aligned} \tag{12}$$

となる。また式 (1)、(2) の $w_u$、$w_\theta$ は

$$\Delta w_u\propto\hat{r}\cdot n_u \tag{13}$$

$$\Delta w_\theta\propto\hat{r}\cdot n_\theta \tag{14}$$

のように更新する。すなわち、もしノイズで振った方向で報酬が得られるならば、その方向に向かって $w_u$、$w_\theta$ を更新するということである。また式 (3)、(4) のノイズ $n_u$、$n_\theta$ の振幅 $N_u$、$N_\theta$ は、初期の振幅を $N_{u0}$、$N_{\theta 0}$ として、

$$N_u=N_{u0}\cdot\exp(-\alpha_u P) \tag{15}$$

$$N_\theta=N_{\theta 0}\cdot\exp(-\alpha_\theta P) \tag{16}$$

のように、$P$ が大きくなるに従って小さくなるようにする。[5]

## 2.2 パラメータの設定について （ボールのシュートの場合）

報酬 $r$ の最大値を $r_{max}=1$ とする。また、初期状態 ($w_u=0$) でランダムに投げる範囲を $u_{max}$ の $a$ 倍であるとすると、

$$a\cdot u_{max}=u_{max}\cdot g(N_{u0})$$

$$N_{u0}=g^{-1}(a) \tag{17}$$

となる。($\theta$ 方向も同様にして、$N_{\theta 0}=\tanh^{-1}(b)$) 最大の報酬が得られている状態でもノイズが $n_{umin}$ だけ残るとすると、

$$n_{umin}=N_{u0}\cdot\exp(-\alpha_u P_{max})$$

$$\alpha_u=\ln\left(\frac{N_{u0}}{n_{umin}}\right)/P_{max} \tag{18}$$

となる。ここで、$P_{max}$ は、$P$ の最大値であり、式 10 において、$\hat{r}=0$、$r=r_{max}$、$P(t)=P(t-1)=P_{max}$ として、

$$P_{max}=\frac{r}{1-\gamma} \tag{19}$$

で算出される。

また、$P(u,\theta)$ を構成する RBF は、$(u,\theta)$ 空間で 10×10 の格子状に並べたものとした。

## 2.3 壁うちのアルゴリズム

ラケットでボールを壁に向かって打ち返し、バウンドしたボールを再び打ち返す、という壁打ちの動作を考える。(図 3) ボールの運動は、粘性抵抗を無視したニュートン方程式に従い、壁や床、ラケットでの反射は、非弾性衝突 (跳ね返り係数 $e=-0.6$) を仮定する。

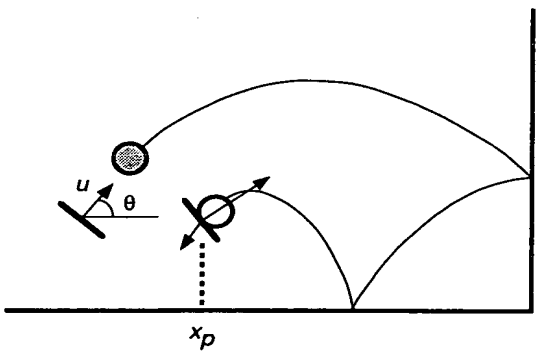


図 3: テニスの壁打ちの概念図

シュートは、予測を必要としない最も簡単な例であったが、予測が必要でかつ環境からの直接のフィードバックがある問題としてテニスの壁打ちを考える。

ラケットの高さは固定とし、プレーヤーが打ち返す位置は水平方向にのみ移動できる。また打ち返すときのラケットの角度と速度を制御できるとする。ボール

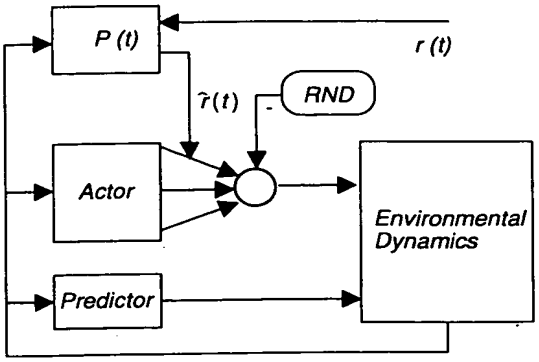


図 4: 壁打ちの予測・制御モデル

が床でバウンドしたときの速さ $v$、角度 $\phi$、位置 $x_b$ から、ラケットとボールの衝突位置 $x_p$ を予測し、ボールを打つ速さ $u$、角度 $\theta$ を制御する。それぞれの式は、

$$u = u_{max} \cdot g(\sum_{ijk} w^u_{ijk} \cdot b_{ijk}(v, \phi, x_b) + n_u) \quad (20)$$

$$\theta = \theta_{max} \cdot g(\sum_{ijk} w^\theta_{ijk} \cdot b_{ijk}(v, \phi, x_b) + n_\theta) \quad (21)$$

$$x_p = x_{max} \cdot g(\sum_{ijk} w^{x_p}_{ijk} \cdot b_{ijk}(v, \phi, x_b) + n_{x_p}) \quad (22)$$

となる。また $b_{ijk}(v, \phi, x_p)$ は、

$$\begin{aligned} b_{ijk}(v, \phi, x_b) &= \exp(-\frac{(v-\mu_i)^2}{2\sigma_v^2}) \\ &\quad \cdot \exp(-\frac{(\phi-\nu_j)^2}{2\sigma_\phi^2}) \\ &\quad \cdot \exp(-\frac{(x_b-\xi_k)^2}{2\sigma_{x_b}^2}) \end{aligned} \quad (23)$$

であり、$P$ の式は、

$$P(t) = \sum_{ijk} w^p_{ijk} \cdot b_{ijk}(v, \phi, x_b) \quad (24)$$

となる。それぞれの荷重値更新の式は以下のようになる。

$$\Delta w^u_{ijk} \propto \hat{r} \cdot n_u \cdot b_{ijk}(v, \phi, x_b) \quad (25)$$

$$\Delta w^\theta_{ijk} \propto \hat{r} \cdot n_\theta \cdot b_{ijk}(v, \phi, x_b) \quad (26)$$

$$\Delta w^{x_p}_{ijk} \propto \hat{r} \cdot n_{x_p} \cdot b_{ijk}(v, \phi, x_b) \quad (27)$$

プレーヤーは、ボールがバウンドした後に移動を開始し、その水平方向の速度 $v_r$ は、ラケットの現在位置 $x_r$ と、予測した衝突地点 $x_p$ に比例した力で加速すると考える。

$$m\frac{dv_r}{dt} = -\eta v_r + a(x_p - x_r) \quad (28)$$

簡単のために、プレーヤーの慣性が、加速力や抵抗力に比べ十分小さいものとすると、移動速度は次式で与えられる。

$$v_r = \frac{a}{\gamma} \cdot (x_p - x_r) \quad (29)$$

以下では、$a/\gamma = 0.5$ とした。正しい予測が行われて、ラケットの高さとボールの高さが一致し、かつラケットの幅以内にボールが当れば、反射則に従って跳ね返すことができるとした。それ以外の場合は、失敗としてサーブからやり直し、以上の動作を繰り返す。

また、$P(t)$ の学習は次のように行う。

$$\Delta w^p_{ijk} \propto \hat{r} \cdot b_{ijk}(v, \phi, x_b) \quad (30)$$

## 3 シミュレーション

### 3.1 シュートの結果

#### 3.1.1 報酬 $r = 1$ とした場合

報酬 $r$ を、ゴールに入ったら $r = 1$、入らなかった場合には $r = 0$ とした。実験で用いたパラメータは、以下の通りである。

このときの報酬 $r(u, \theta)$ の様子を図 5に示す。直接ゴールに入る場合や壁に反射して入る場合、一旦床に反射して入る場合など多数の解が共存しているのが分かる。また図 6に過去１００回あたりの成功率を、図 7に学習後の $P(u, \theta)$ の形を示す。収束する位置は乱数の系列によって違う。

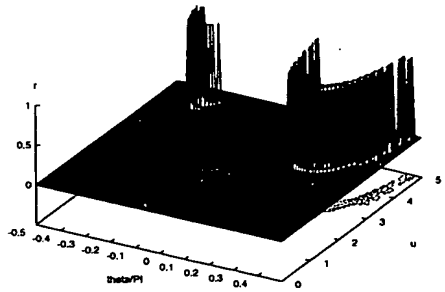

図 5: 報酬 $r(u, \theta) = 1$ の様子

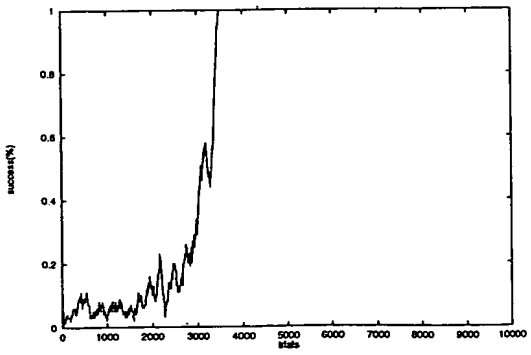

図 6: 過去１００回の成功率 ($r = 1$ の場合)

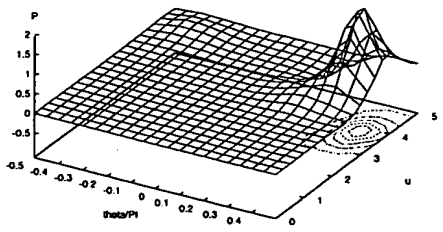

図 7: $r = 1$ のときの $P(u, \theta)$

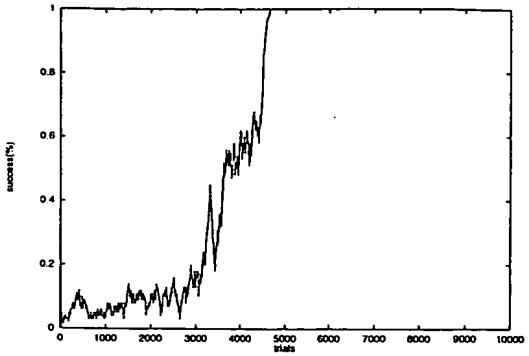

図 9: 成功率 (報酬 $r$ に傾斜をつけた場合)

### 3.1.2 報酬 $r$ に傾斜をつけた場合。

報酬に傾斜をつけることで、最適な $(u, \theta)$ で投げることを学習するかどうかを調べた。ゴールに入る解のうち、ボールがゴールの中心を通りしかも最小の初期速度の場合の速度を $u_{min}$ とする。この速度 $u_{min}$ が最も遅い速度であり、この速度で打てば、エネルギーを最も使わずに、より高い報酬が得られると考え、報酬 $r$ を

$$r = r_{max} - \frac{(u - u_{min})^2}{(u_{max} - u_{min})^2} \tag{31}$$

とした。報酬 $r(u, \theta)$ の様子を図8に示す。先の $r = 1$ としたものに傾斜がついたものであることがわかる。$u_{min}$ は、計算により、$u_{min} = 2.714$ である。図9に成功率を、学習後の $P(u, \theta)$ を図10に示す。最終的に $u = 3.60, \theta = 0.38\pi$ 付近に収束している。これは最適な $u$ の値 $u_{min} = 2.71$ にはなっていない。傾斜が緩いため、ある程度の報酬が得られると、ノイズがそれに比例して小さくなり、そのため途中で止まっていると考えられる。

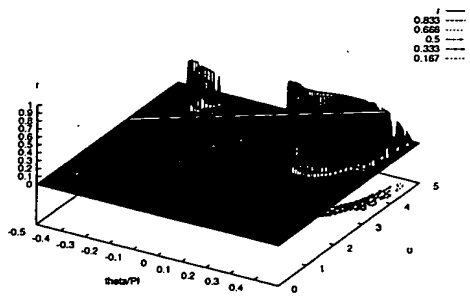

図 8: $r = r_{max} - \frac{(u-u_{min})^2}{(u_{max}-u_{min})^2}$

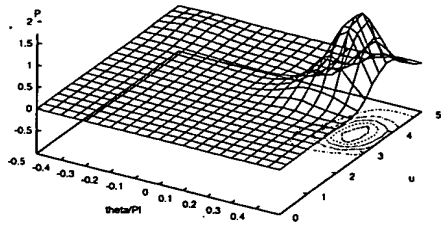

図 10: $P(u, \theta)$ (報酬 $r$ に傾斜をつけた場合)

### 3.1.3 報酬 $r$ の傾斜を変えた場合

次に、報酬 $r$ の傾斜を変えて、最適な $(u, \theta)$ に行くかどうか調べた。報酬 $r$ を

$$r = \frac{u_{max} - u}{u_{max} - u_{min}} \leq r_{max} \tag{32}$$

とした。(図11) この場合、図12に示すように全く成功していない。ランダムに投げる場合の中心点、$u = u_{max} \cdot g(w_u), \theta = \theta_{max} \cdot \tanh(w_\theta)$ の動きを図13に示す。$u, \theta$ は $\pm u_{max}, \theta_{max}$ 付近を中心としていることがわかる。図14の $P(u, \theta)$ をみると、$P$ に負の部分ができている。その部分に投げたことで、$P$ が負となり、$w_u, w_\theta$ が大きな値になったため、中心点が $u, \theta$ の境界付近になってしまったと考えられる。

### 3.1.4 $v_{ij} \geq 0$ に制限した場合

そこで、$P$ の荷重値 $v_{ij}$ を $v_{ij} \geq 0$ に制限することで、$P$ に負の部分ができないようにした。成功率を図15に、学習後の $P(u, \theta)$ の様子を図16に示す。このときの $u$ は最適ではないが、報酬の得られる上の部分では最も報酬が大きくなる値になっている。

## 3.2 壁うちの結果

テニスの壁打ち動作において、報酬は、連続して打ち返すことが出来た場合に与えられると考えるのが

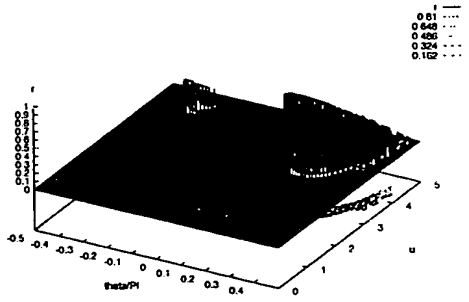

図 11: $r = \frac{u_{max}-u}{u_{max}-u_{min}} \leq r_{max}$

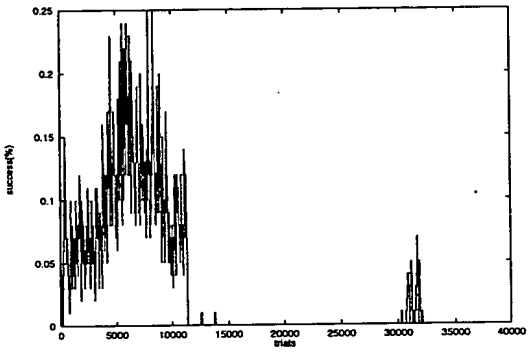

図 12: 報酬の傾斜を変えた場合の成功率

妥当である。しかし、シュートと同じように壁うちの場合も可能な解は無数にあるため、可能な解の間に何らかの優劣を与える必要がある。あまり壁に近くてショートバウンドで打ち返すのも得策ではないし、壁から離れすぎては、大きなエネルギーが必要になる。また、打ち返す位置が毎回変動するのも好ましくない。そこで、ここではなるべく同じ位置で打ち返すように報酬をラケットの位置 $x_p$ にたいして、

$$r = \exp\left(-\frac{(x_p - x_0)^2}{2\sigma^2}\right) \qquad (33)$$

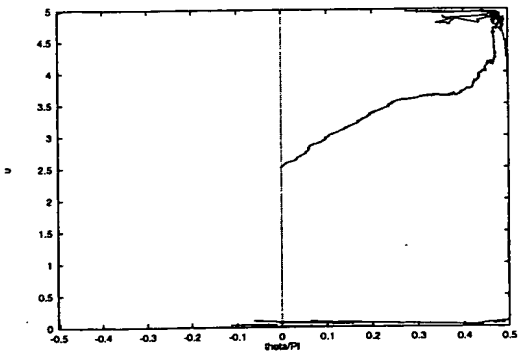

図 13: 中心点の動き

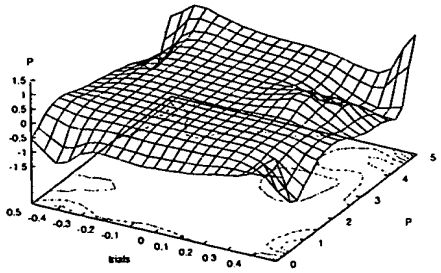

図 14: $P(u,\theta)$ （報酬の傾斜を変えた場合）

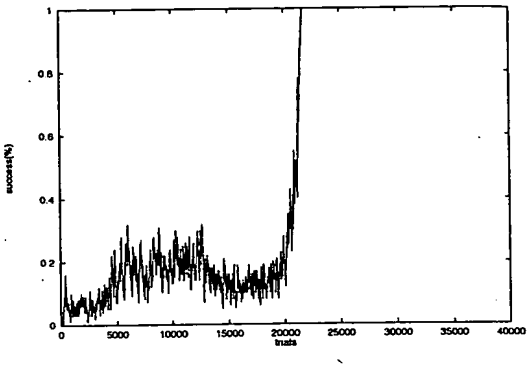

図 15: 成功率（$v_{ij} \geq 0$ の場合）

で与えた。$x_0 = 6, \sigma = 1$ とした。また報酬は打ち返すことが出来た場合にのみ与えるようにした。図17に連続して打ち返した回数を示す。３０００回ぐらいで連続して打ち返せるようになっている。このとき打ち返す位置 $x_p$ （図18）をみると、$x_0$ に近い値になっている。

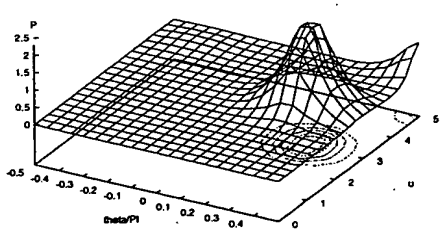

図 16: $P(u,\theta)$ （$v_{ij} \geq 0$ の場合）

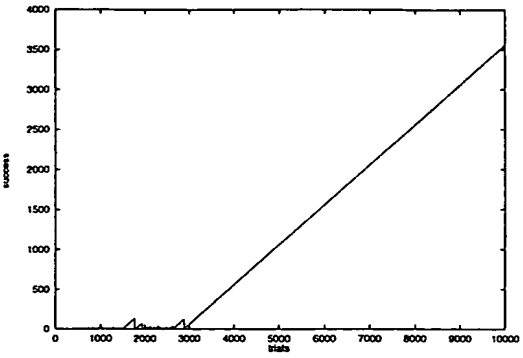


図 17: 連続成功回数

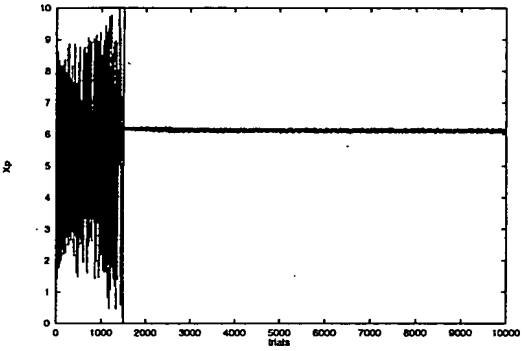


図 18: 衝突位置の予測 $x_p$

## 4 まとめと考察

### 4.1 シュート

最も簡単な課題としてボールのシュートを考え、アクティブな探索や報酬による強化学習、RBF による学習の高速化が有効に働く事を確認した。特に、外界の知識が得られない場合には、ノイズを用いた確率的強化学習が有効である。また、この問題では、可能な解が無数に存在するため、オンライン学習では、初期値や乱数により異なった解に収束する。

解に優劣を付けた実験では、報酬 $r$ に傾斜を付けた場合、傾斜が緩やかであると最適な値に行く前に収束してしまう場合がある。傾斜を変えた場合には、$P$ に負の部分ができ、収束しないことがある。これは、$P$ の荷重値を正に制限することで、最適な値に近付けることができた。

### 4.2 壁打ち

テニスの壁打ちの課題については、ボールの運動法則を学習し、落下地点を予測するとともに、望ましい位置で連続して打ち返せるようラケットの角度と速度を制御できるようになった。ここで用いたボールの運動予測 (外界の内部モデル) は、ニュートン方程式を解いているわけではないので仮に空気抵抗が未知であったり、ボールや壁の跳ね返り係数を知らなくても対応できる点にメリットがあり、複雑で動的な環境に適応できる可能性を持っている。同じことは、ラケットの制御アルゴリズムについても言える。しかし、以上のシミュレーションにおいては、ラケットにボールが当るタイミングに関する制御は行っていない。実際のロボット制御などに応用するためには、ラケットを加速する時間が必要なため、一旦、take back した後に、衝突のタイミングを予測して打ち返すことが要求される。今後は、このようなタイミングの予測・制御と連続的な動作を含んだアルゴリズムについて検討を行う。

## 参考文献

[1] A. G. Barto, R. S. Sutton, and C. W. Anderson, IEEE Transaction on System, Man, and Cybernetics, SMC-13 (1983) 834-846.

[2] R. S. Sutton, Machine Learning,**3** (1988) 9-44.

[3] A. G. Barto, in *Models of Information Processing in the Basal Ganglia*, ed. by J. C. Houk, J. L. Davis, and D. G. Beiser, (1995, MIT Press) pp. 215-232.

[4] K. Doya, in *Advances in Neural Information Processing Systems 8*, ed by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, (1996, MIT Press)

[5] V. Gullapalli, Neural Networks, 3 (1990) 671-692.

(盛 文 社)